DOCUMENT RESUME

ED 283 872 TM 870 410

AUTHOR Williams, Janice E.
TITLE Sample Size in Analyzing Regressions across Two Groups.
PUB DATE Apr 87
NOTE 15p.; Paper presented at the Annual Meeting of the Western Psychological Association (67th, Long Beach, CA, April 23-26, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Simulation; *Estimation (Mathematics); *Monte Carlo Methods; *Quasiexperimental Design; *Regression (Statistics); *Sample Size; Sampling; Statistical Distributions; Statistical Studies
IDENTIFIERS Random Numbers

ABSTRACT

A Monte Carlo study was done to determine the adequate sample size for quasi-experimental regression studies, which compare regression lines for two groups and estimate their point of intersection. Populations of 1,000 subjects in each of two groups were constructed (using random normal deviates) to yield equivalent regression lines of opposite slope. Multiple runs of subset samples of 20, 25, 30, 40, and 50 points per group provided the sampling error of estimating X- and Y-coordinates for the intersection point. Standard error was an inverse function of group size (degrees of freedom), as with other sampling statistics. The function leveled off at about 50 subjects per group, suggesting this as the recommended minimal group size for such analyses. The author notes that the acceptable sample size for estimating the intersection of two lines (N=50/group) exceeds that for estimating the difference between sample means (N=30/group) because of the additional parameter involved. (LPG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED283872

SAMPLE SIZE IN ANALYZING REGRESSIONS ACROSS TWO GROUPS

Janice E. Williams
Graduate School of Education
Research Methods and Evaluation
U.C.L.A.

Paper presented to the Western Psychological Association,
Long Beach, California on April 23, 1987

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J.E. Williams

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☒ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 870410

SAMPLE SIZE IN ANALYZING REGRESSIONS ACROSS TWO GROUPS

Janice E. Williams
Graduate School of Education
Research Methods and Evaluation
U.C.L.A.

When analyzing regressions across groups, the slopes should be compared and the regression lines checked for a possible intersection (the region of interest). Deviations from the point of intersection may be used to make conclusions about significant differences between treatment groups. However, the point of intersection may fluctuate depending upon sample sizes, means, and residual mean squares.

A Monte Carlo study was conducted at various levels of sample size to determine the expected variability of the coordinates of the point of intersection. Results from this research yielded a reasonable indication of the sampling error due to sample size alone. The relative magnitude of the standard error of the point of intersection may be an inverse function of the total degrees of freedom. Given the parameters of this study, an investigator would be wise to select groups with N=50 or more in order to minimize error variability in estimates of the coordinates of the point of intersection.

SAMPLE SIZE IN ANALYZING REGRESSIONS ACROSS TWO GROUPS

Janice E. Williams
Graduate School of Education
Research Methods and Evaluation
U.C.L.A.

A paper presented to the Western Psychological Association,
Long Beach, California on April 23, 1987

Quasi-experimental research is very common. Such designs are used extensively in field settings (Campbell & Stanley, 1966; Boruch, Wortman & Cordray, 1981). Although there is no general agreement on the appropriate method of obtaining unbiased statistical estimates of treatment effect (Cronbach, et al., 1977; Lord, 1967), regression is generally used to analyze these designs. A strategy of multiple tests has been emphasized to bracket the true effect of treatment (Wortman, Reichardt & St. Pierre, 1978). The resulting conclusions are then robust over a range of conditions.

When analyzing regressions across groups, the slopes should be compared for each group and the regression lines checked for a possible intersection (the region of interest). Deviations from the point of intersection may then be used to make conclusions about significant differences between two treatments, or two groups at any specific point of a continuous variable X (Pedhazur, 1982). Areas where the expected performance differs for the two groups represented by a categorical variable may thus be compared. However, the point of intersection may fluctuate

depending upon sample sizes, means, and residual mean squares.

There are few published studies on the parameters of the point of intersection of two different regressions. Possibly this is a consequence of the difficulty of mathematically determining the sampling distribution of the point of intersection. Such an equation might appear to involve eight different parameters:

For group 1; Rxy, $\bar{X}$, N, Residual M.S.

For group 2; Rxy, $\bar{X}$, N, Residual M.S.

As a consequence of this lack of information, no realistic guidelines have been cited relating sample size per group to the expected variability in the point of intersection. Without such information, an investigator is left in a quandry concerning appropriate group size: Should the total N be 30, 100, 300 or ?? Without such critical information, quasi-experimentation is simply a matter of guess-work.

This problem appears to be nearly mathematically intractable (or at least not easily solved). Thus a Monte Carlo study at various levels of sample size (e.g. 20, 25, 30, 40, 50) would appear to yield a reasonable indication of the sampling error involved in such a process. Such data should serve as a guide to the design of quasi-experimental research involving two groups.

DIMENSIONS OF THE PROBLEM

A microcomputer was used to generate X, Y, Z, and XZ scores for a population of 2000 computer subjects, divided into two groups of 1000 subjects each. Two series of random normal numbers were drawn using the random number generator of an Epson

QX-10, and transformed to normal deviates (p to z). An a priori decision led to correlated X and Y scores for each group. A constant of 1.0 was added to each Y to offset the distribution. This offset was one standard deviation on the Y due to the p to z transformation.

The X and Y scores for Group 1 were generated with the following equation: Y = .7071 X + .7071 Random normal deviate + 1, while the equation: Y = -.7071 X + 0.7071 Random normal deviate generated the X and Y scores for Group 2. A pictorial representation of the theoretical situation is shown in Figure 1. A model was then constructed from Figure 1 to allow for a

---

Insert Figure 1 about here

---

comparison of the regression lines for two groups. In addition, variable Z scores were generated for dichotomous groups, which were dummy coded 0 or 1, as was the XZ interaction term. The theoretical point of intersection for the two groups was toward the center of the graph shown in Figure 1.

METHOD

All analyses for the current study were run using the GUYL STATPAK (Micro/Research, 1983). Normal distribution parameters for the entire population of 2000 were then computed. These parameters closely corresponded to the theoretical model which had been constructed. The computed values indicated that the offsets were approximately correct. Table 1 reveals that all standard deviations were approximately equal to a sigma of one, and skewness and kurtosis were approximately normal. Pearson

correlations (Table 1) indicated that the matrices for the two

---

Insert Table 1 about here

---

groups also corresponded to the theoretical model constructed. As a consequence, given the structure of the design, the underlying assumptions of the multivariate normal distribution and linearity of regression may be assumed for the population. Any sample from this population may show some deviation, but such values would be comprised of random deviation due to sampling error - not true differences.

Once this general population of scores had been tested against the model, subset samples (N=20, 25, 30, 40 ,50) were drawn from the sample population. The number of computer runs (data points) were 50 for N=20, 40 for N=25, 33 for N=30, 25 for N=40, and 20 for N=50. Points of intersection were then computed for each sample: the X and Y coordinates. The standard deviation of these X and Y coordinates provide the primary data of the current study: the standard error of the X-intercept and the standard error of the Y-intercept, denoted as σ_{Ix} and as σ_{Iy}, respectively.

RESULTS

The resulting standard errors, as a function of increasing sample size, are illustrated in Figure 2 for σ_{Ix} and in Figure 3 for σ_{Iy}. It should be noted that the obtained values for σ_{Iy} are consistently less than the obtained values for σ_{Ix}. This result was anticipated since the regression direction for each line was computed as $\hat{Y} = a + bX$. Each predicted Y value

contained only a portion (b^2) of the variability of the corresponding X values. In the present example, the σ_{Iy} = 0.7477 compared to an expected relationship of the σ_{Iy} = 0.7071 due to the symmetrical r = 0.7071 for each of the two lines.

As expected, Figures 2 and 3 show a consistent decline in

Insert Figures 2 and 3 about here

---

the magnitude of the standard error of the intercept as some inverse function of group size. No attempt was made to determine the exact equation since only equal size groups were used for this illustration. It is, however, reasonable to assume that these standard errors should be some inverse function of total d.f. for the two groups. Other statistical tools (F, t, T^2, etc.) also exhibit this inverse relationship between the magnitude of the critical statistic and total d.f. in the sample. Analogies to the current investigation include the standard error of a sample mean, the standard error of the difference between two sample means, the standard error of a b-weight, etc.: all inverse functions of d.f. for sample size. Total d.f. was assumed to be [(N1-3) + (N2-3)] since three parameters had to be determined for each group: mean, standard deviation and correlation coefficient. The resulting correlations between sample size and σ_I were 0.969 for the X-intercept standard error and 0.956 for the Y-intercept standard error.

DISCUSSION

The Monte Carlo results indicate that the relative magnitude of the standard error of the point of intersection (σ_I)

may be an inverse function of the total degrees of freedom [(N1-3)+(N2-3)]. This function is graphically plotted as Figure 4. In that figure, the function appears to level off (or become

---

Insert Figure 4 about here

---

asymptotic) somewhere near N=50 per group. Thus an investigator would be wise to select groups with N=50 or more in order to minimize error variability in estimates of the coordinates of the point of intersection. There is parsimony in knowing the appropriate number of subjects to run in a non-experimental research setting. It should be noted that N/group=50 is somewhat higher than corresponding levels for standard errors of the difference between two sample means, where N/group=30 to approach asymptoticity. This difference in optimal group size may be due to the difference in the number of parameters being investigated: 3 parameters must be known for each group for point-of-intersection estimates but only 2 parameters per group must be known for differences-between-means estimates.

It should be apparent that the current study has investigated only a single condition of the variability of the coordinates of the point of intersection. It would be interesting to conduct a full Monte Carlo study of these coordinates varying the number of subjects per group (N1=20, N2=40; etc.), the relative magnitude of the slopes of the regression lines (r_1 =.8, r_2 =.3; etc., and the within-group variability (σ_1 =1, σ_2 =3; etc.). With such a study, an estimate of the function of the value of σ_I could be determined.

REFERENCES

Boruch, R.F., Wortman, P.M. & Cordray, D.S. (1981). Reanalyzing Program Evaluations. San Francisco: Jossey-Bass Publ.

Campbell, D.T. & Stanley, J. (1966). Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally.

Cook, T.D. & Reichardt, C.S. (1976). Guidelines- statistical analysis of non-equivalent control group designs: a guide to some current literature. Evaluation 3, 136-138.

Cronbach, L.J., Rogosa, D.R., Floden, R.E. & Price, G.G. (1977) Analysis of covariance in nonrandomized experiments: parameters affecting bias. Occasional Paper, Stanford Evaluation Consortium, Stanford University (February).

Lord, F.M. (1967). A paradox in the interpretation of group comparisons. Psychological Bulletin 68, 304-305.

Micro/Research (1983). Guyl Statpak Forest Ranch, CA.

Pedhazur, E.J. (1982). Multiple Regression in Behavioral Research. New York: Holt, Rinehart & Winston.

Wortman, P.M., Reichardt, C.S. & St. Pierre, R.G. (1978). The first year of the education voucher demonstration. Evaluation Quarterly 2, 193-215.

Table 1

Theoretical Parameters and Correlation

Variable	Mean	Sigma	Skewness	Kurtosis	Correlation
X-Coordinate	0	1	0	3	$\lvert .7071 \rvert$
Y-Coordinate	*	1	0	3	

*Note: $\bar{Y}_T = 1$ for group 1; $\bar{Y}_T = 0$ for group 2

Normal Distribution Parameters and Correlations

Group 1

Variable	Mean	Sigma	Skewness	Kurtosis	Correlation
X-Coordinate	.9876	1.0145	.02189	2.921	.707175
Y-Coordinate	-.0187	1.0350	-.07104	3.052	

Group 2

Variable	Mean	Sigma	Skewness	Kurtosis	Correlation
X-Coordinate	-.0745	1.0199	-.10401	2.884	-.702561
Y-Coordinate	-.0597	1.0202	.03622	2.964	

Actual (X,Y) and Theoretical (X_T, Y_T) Point of Intersection

X, Y	-0.72925, 0.49507
X_T, Y_T	-0.70711, 0.50000

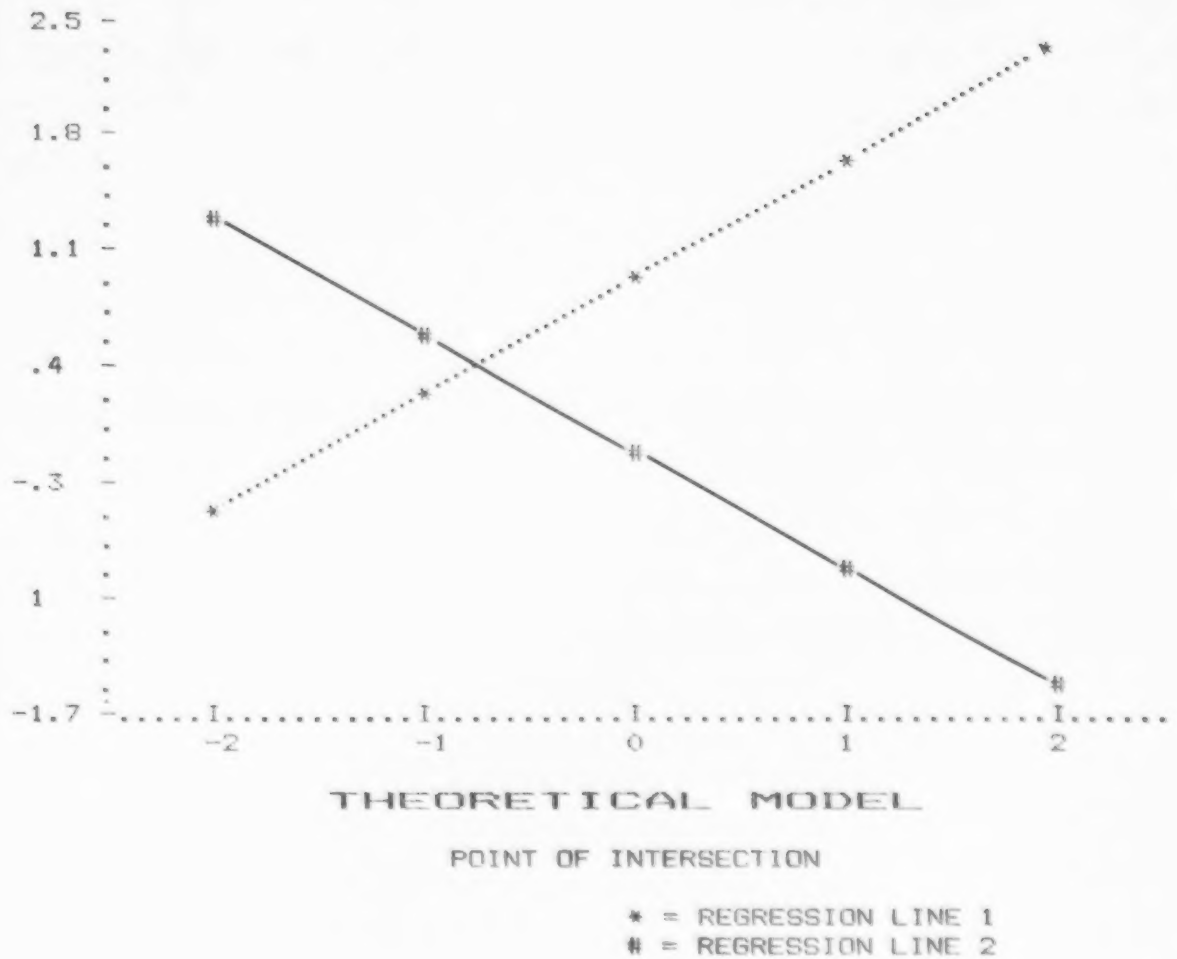


Figure 1. Theoretical Model: Point of Intersection

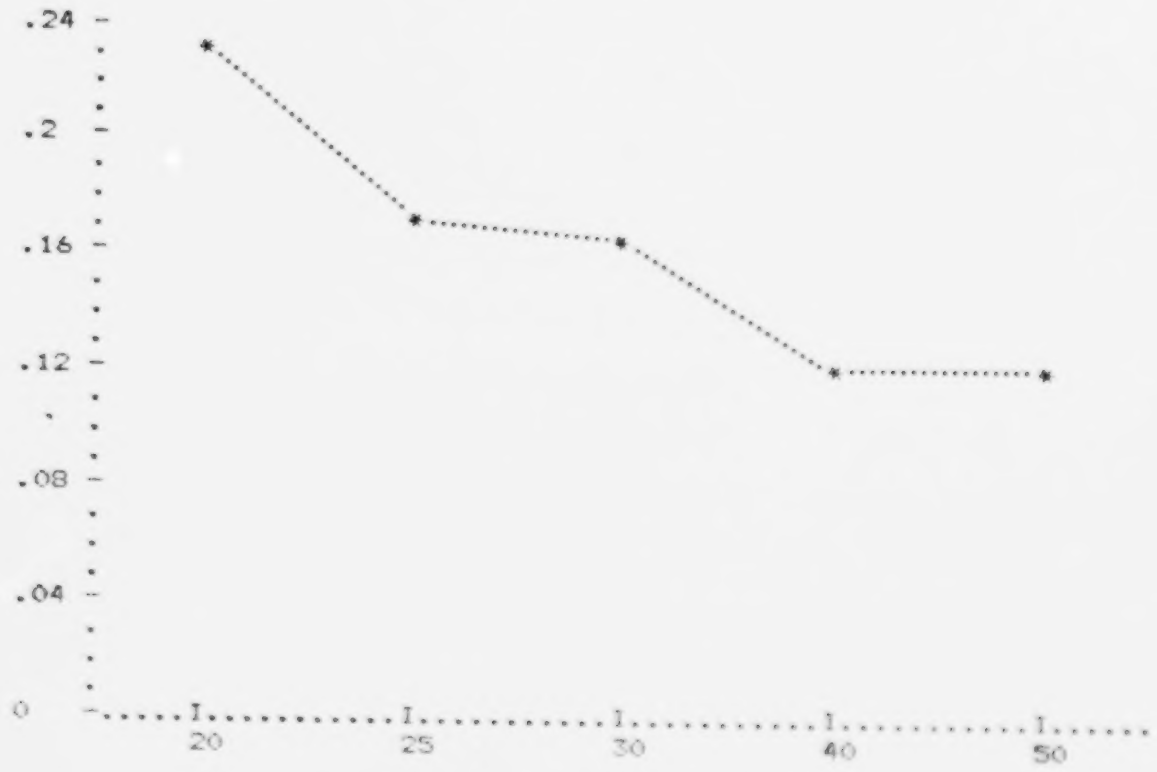


Figure 2. Standard Error of Point of Intersection for X

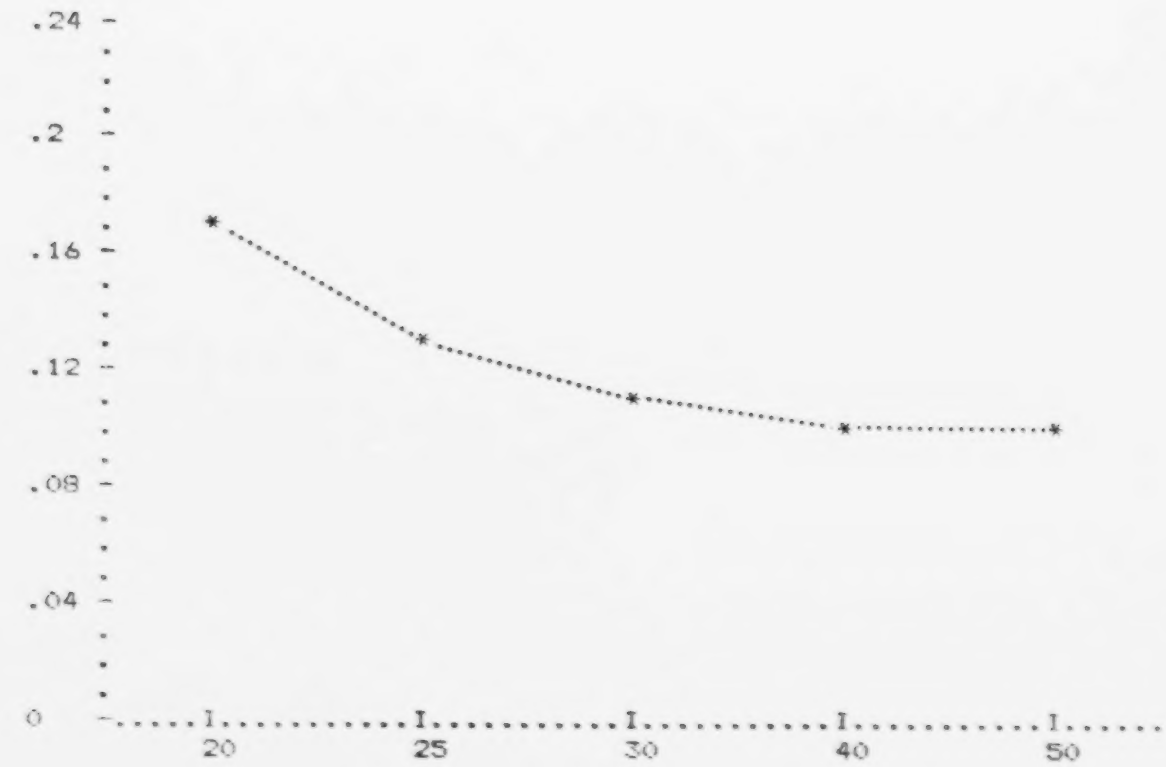
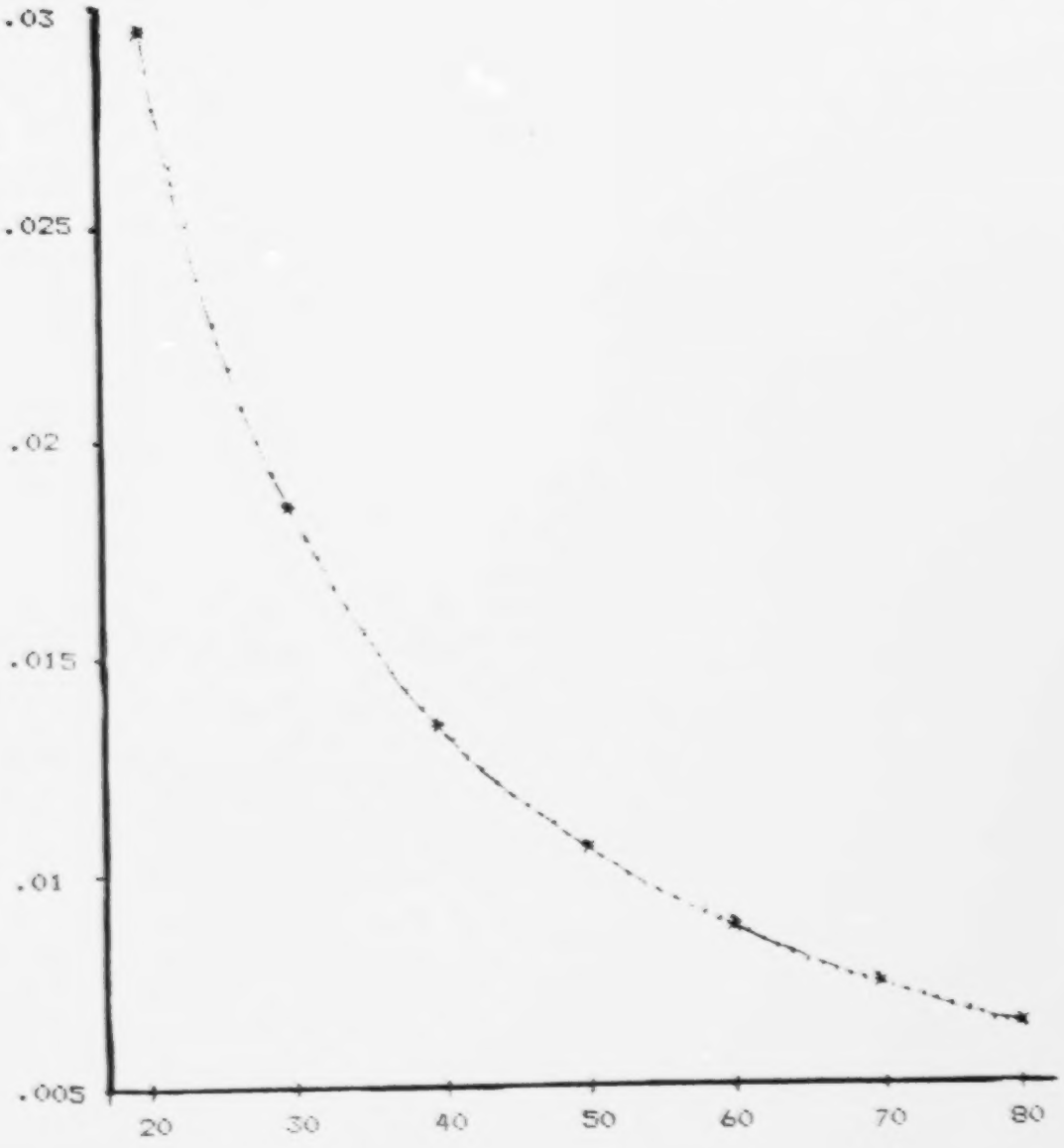


Figure 3. Standard Error of Point of Intersection for Y



RELATIVE ERROR

FUNCTION = 1/[(N1-3)+(N2-3)]

KEY:
20 = N PER GROUP
30 = N PER GROUP
40 = N PER GROUP
50 = N PER GROUP
60 = N PER GROUP
70 = N PER GROUP
80 = N PER GROUP

* = RELATIVE ERROR

Figure 4. Relative Error Related to Total Degrees of Freedom